DOCUMENT RESUME

ED 380 502 TM 022 865

AUTHOR Longford, Nicholas T.
TITLE A Case for Adjusting Subjectively Rated Scores in the Advanced Placement Tests. Program Statistics Research. Technical Report No. 94-5.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
REPORT NO ETS-RR-94-58
PUB DATE 94
NOTE 25p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Advanced Placement; Computer Science; English; Error of Measurement; *Essay Tests; Grading; High Schools; High School Students; Psychology; *Reliability; Scores; *Scoring; *Student Placement; *Writing Tests
IDENTIFIERS Advanced Placement Examinations (CEEB); *Subjectivity

ABSTRACT

A case is presented for adjusting the scores for free response items in the Advanced Placement (AP) tests. Using information about the rating process from the reliability studies, administrations of the AP test for three subject areas, psychology, computer science, and English language and composition, are analyzed. In the reliability studies, 299 psychology essays, 250 and 248 essays for two forms of the computer science examination, and 250 essays for English were rated. It is shown that the minimum squared error score adjustments proposed by N. T. Longford (1993) result in changed AP grades for an appreciable percentage of examinees. The proposed schemes are easy to implement and involve no iterative procedures. Four tables and two figures present details of the analyses. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-94-58

A Case for Adjusting Subjectively Rated Scores in the Advanced Placement Tests

Nicholas T. Longford
Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 380 502

PROGRAM STATISTICS RESEARCH

Technical Report No. 94-5

Educational Testing Service
Princeton, New Jersey 08541

TM022845

Copyright © 1994. Educational Testing Service. All rights reserved.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

A case for adjusting subjectively rated scores in the Advanced Placement tests *

Nicholas T. Longford

Abstract

The purpose of this report is to present a case for adjusting the scores for free response items in the Advanced Placement tests. Using information about the rating process from the reliability studies, administrations of the Advanced Placement test for three subject areas are analyzed, and it is shown that the minimum squared error score adjustments proposed by Longford (1993) result in changed AP grades for an appreciable percentage of the examinees.

Key words: mean squared error; score adjustment; shrinkage; variance components.

*Data for the analyses described in this report were provided by Behroz Maneckshana and Rick Morgan. Minhwei Wang's assistance with computing and data handling is acknowledged. Reviews by Nancy Petersen, Walter MacDonald, and Rick Morgan contributed to improvements on the earlier versions of the manuscript. Research for this report was funded by the Program Research Planning Council.

1 Introduction

Testing programs using constructed response items have had to face the dilemma whether to adjust the scores for rater severity ever since the first indications that raters may vary in their severities. Longford (1993) introduced several adjustment schemes which make use of the information obtained from a random-effects analysis of variance. Given the information about the variance components due to the true scores, severity, and inconsistency, the resulting adjustments yield scores that are better than the unadjusted or the fully adjusted scores. The adjustment schemes can be applied even when each essay is rated only once, so long there is some external information about inconsistency variation.

The improvement of the mean squared errors for the adjusted scores over their counterparts for the unadjusted scores provides only a very indirect measure of the impact of the adjustment. The impact is more directly assessed by the percentage of students whose aggregate scores change as a result of the adjustment. In the Advanced Placement (AP) tests a simple measure of the impact is defined as the percentage of students whose AP grade would be changed as a result of the adjustment.

In the study reported here we reanalyzed the reliability studies for AP Psychology (Form 3PBP), Computer Science (Form 3PBP), Computer Science (Form A–3PBP), and English Language and Composition (Form 3PBP), and used the results of these analyses to adjust the scores of the corresponding operational administrations. The adjusted scores were applied to determine the 'adjusted' AP grades and the proportion of examinees whose score would be altered as a result of the adjustment. The results are summarized in Table 1 for three adjustment schemes. Details of the adjustment schemes are given in Section 3.

INSERT TABLE 1 HERE

The study demonstrates that a substantial proportion of the AP grades would be altered if either score adjustment scheme were applied. The impact of each score adjustment scheme is stronger when the essays are given greater rel-

ative weight in the composite score, when the inconsistency variation is greater, and, with the exception of the scheme **tAdj**, when the variation in rater severity is greater. Each of the studied score adjustment schemes can be applied when most essays are rated only once, but rerating of a random sample of the essays (as done in reliability studies) is necessary, unless there is another source of information about the variance components.

Section 2 summarizes the variance component models for rater reliability and Section 3 gives details of the adjustment schemes. A more comprehensive background to the methods applied can be found in Longford (1993). Section 4 describes the datasets from the analyzed administrations and reliability studies of the Advanced Placement tests and gives details of the analyses of these datasets. In brief, the adjustment schemes are applied for each essay topic, and the AP grades are recalculated using the adjusted scores. Section 5 contains the conclusions of the study and recommendations for conducting future reliability studies. The recommendations echo the findings of Braun (1988), although the proposed procedures can be implemented without any changes in the design of the reliability studies.

2 Models

We consider models for two kinds of rating exercises. In a *reliability study* each essay is rated by two distinct raters. The rating in the first session is referred to as the *operational* rating and the rating in the second session as the *experimental* rating. Differences between the two ratings of an essay can be attributed to the amalgam of the differences in severity of the raters, disagreement in merit, and the temporal variation in the raters' decisions. These influences are represented by the additive model

$$y_{i,j,k} = \alpha_i + \beta_{j,k} + \varepsilon_{i,j,k}\,, \tag{1}$$

where $y_{i,j}$ denotes the score given by rater $j = 1, 2, \ldots, J$ to essay/examinee $i = 1, 2, \ldots, I$, α_i is the true score of examinee i, β_j is the severity of rater j, and $\varepsilon_{i,j}$ is a residual term representing inconsistency. The complex subscript

notation $(j,_k)$ is used to denote the rater who graded essay i in session $k = 1, 2$. Thus all the (realized) scores given by the raters are $y_{i,j,_k}$, $i = 1, \ldots, I$ and $k = 1, 2$. Further, denote by n_j the number of essays graded by rater j, and by N the total number of ratings, $N = 2I = n_1 + \ldots + n_J$. We refer to n_j as the workload of rater j. Throughout, it is assumed that the allocation of essays to raters is non-informative.

The terms α_i, β_j, and $\varepsilon_{i,j}$ are assumed to be independent random variables with means μ, 0, and 0, and variances σ_a^2 (due to the true scores), σ_b^2 (due to rater severity), and σ_e^2 (due to inconsistency), respectively.

When each essay is rated only once, the subscript $k = 1$ is redundant. With no replication of the rating of an essay the variation in true scores cannot be disentangled from the inconsistency variation. However, information about inconsistency variation can be obtained by rerating a random sample of the essays using raters selected from the same population of raters as for the first (operational) scoring. Reliability studies, conducted for the AP tests on a regular schedule, can be effectively used for this purpose.

We apply the following procedure for combining information from reliability studies and operational scoring: First, the reliability study scores are analyzed to obtain the estimates of the variance components σ_a^2, σ_b^2, and σ_e^2. Next, the operational scores are analyzed using the estimate of the inconsistency variance σ_e^2 from the reliability study. Methods for estimating the variance components are described in the next section. For full details and extensions, see Longford (1993).

The variance components play an important role in the score adjustment schemes described in Section 3. To motivate these schemes consider the following extreme scenarios. If we knew that the raters did not differ in severity, any adjustment for severity would be counterproductive. If the examinees did not vary in their true scores each examinee should be given the same score, even though the raters have given scores on the entire score scale. A realistic situation contains elements of each extreme scenario. This suggests shrinkage estimation of the true scores as a compromise between the two extremes. The scores should

be pulled closer to the overall mean when the raters wildly disagree and they should be adjusted for severity when the raters' severities are well determined.

2.1 Estimation

This section gives details of the moment matching method for estimating the variance components when each essay is rated twice (that is, in two complete sessions). The variance components are estimated separately for each constructed response item.

Let

$$y_{i,\cdot} = \frac{y_{i,j_{i1}} + y_{i,j_{i2}}}{2}$$

be the unadjusted mean essay score for examinee i graded in a reliability study,

$$z_j = \frac{1}{n_j} \sum_{k=1}^{2} \sum_{i:j_{ik}=j} y_{i,j}$$

be the mean score given by rater j, and

$$\bar{y} = \frac{1}{I} \sum_{i=1}^{I} y_{i,\cdot}$$

be the mean of the observed scores on the essay. We define the following statistics:

S_E, the within-essay sum of squares,

$$S_E = 2 \sum_{i=1}^{I} (y_{i,j_{i1}} - y_{i,j_{i2}})^2;$$

S_R, the within-raters sum of squares,

$$S_R = \sum_{k=1}^{2} \sum_{i=1}^{I} (y_{i,j_{ik}} - z_{j_{ik}})^2;$$

and S_T, the total sum of squares,

$$S_T = \sum_{k=1}^{2} \sum_{i=1}^{I} (y_{i,j_{ik}} - \bar{y})^2.$$

The moment matching equations are

$$\begin{aligned} I(\hat{\sigma}_b^2 + \hat{\sigma}_e^2) &= S_E \\ (2I - J)(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) &= S_R \\ (2I - 1)(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) + \left(2I - n^{(2)}\right)\hat{\sigma}_b^2 &= S_T\,, \end{aligned} \tag{2}$$

where $n^{(2)} = \sum_j n_j^2/(2I)$ is the normalized sum of squares of the raters' workloads. The solution of this system of linear equations in the variance estimates $\hat{\sigma}_a^2$, $\hat{\sigma}_b^2$, and $\hat{\sigma}_e^2$ is

$$\begin{aligned} \hat{\sigma}_b^2 &= \frac{S_T - \frac{2I-1}{2I-J}S_R}{2I - n^{(2)}} \\ \hat{\sigma}_e^2 &= \frac{S_E}{I} - \hat{\sigma}_b^2 \\ \hat{\sigma}_a^2 &= \frac{S_R}{2I - J} - \hat{\sigma}_e^2\,. \end{aligned} \tag{3}$$

The inconsistency variance σ_e^2 cannot be identified when each essay is read only once (as in operation), and so its estimate from the reliability study (double reading) is substituted in the analysis of the operational data. When each essay is scored only once, the within-raters and total sum of squares are

$$\begin{aligned} S_R &= \sum_{i=1}^{I}(y_{i,j_i} - z_{j_i})^2; \\ S_T &= \sum_{i=1}^{I}(y_{i,j_i} - \bar{y})^2 \end{aligned} \tag{4}$$

(omitting the session-subscript $k = 1$), and the within-essay sum of squares is not defined. Assuming an estimate of the inconsistency variance, $\hat{\sigma}_e^2$, the moment matching equations for the other two variances are

$$\begin{aligned} (I - J)(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) &= S_R \\ (I - 1)(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) + (I - n^{(2)})\hat{\sigma}_b^2 &= S_T\,. \end{aligned}$$

The solution of this system of two linear equations is

$$\begin{aligned} \hat{\sigma}_a^2 &= \frac{S_R}{I-J} - \hat{\sigma}_e^2 \\ \hat{\sigma}_b^2 &= \frac{S_T - (I-1)(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)}{I - n^{(2)}} . \end{aligned} \tag{5}$$

In the application, we analyze the operational and experimental scores as two sessions in the reliability study, and then impute the estimate of the inconsistency variance in the analysis of the operational scores. An additional purpose of this approach is to demonstrate how external information about the variance components could be used for score adjustment.

3 Adjustment schemes

Traditionally, adjustment of scores has been considered as a form of compensation for the differences in severity of the raters. This motivates the scheme based on the estimated severity of each rater. A natural estimate of the severity is the difference of the rater's mean score from the mean score in the entire rating exercise:

$$b_j = z_j - \tilde{y} = \frac{1}{n_j} \sum_{(i,\, j_i = j)} y_{i,j_i} - \frac{1}{I} \sum_i y_{i,j_i} .$$

As demonstrated in Longford (1993) this estimator can be improved by shrinking it toward zero.

$$\hat{\beta}_j = s_j b_j .$$

The shrinkage coefficient s_j is chosen so as to minimize the mean squared error $\mathbf{E}(\hat{\beta}_j - \beta_j)^2 = C_{j,0} - 2C_{j,1}s_j + C_{j,2}s_j^2$, where

$$\begin{aligned} C_{j,0} &= \sigma_b^2 \\ C_{j,1} &= \sigma_b^2 \left(1 - \frac{n_j}{I}\right) \\ C_{j,2} &= \sigma_a^2 \left(\frac{1}{n_j} - \frac{1}{I}\right) + \sigma_b^2 \left(1 - \frac{2n_j}{I} + \frac{n^{(2)}}{I}\right) + \sigma_e^2 \left(\frac{1}{n_j} - \frac{1}{I}\right) \end{aligned} \tag{6}$$

($n^{(2)} = \sum_j n_j^2/I$), as a function of the shrinkage coefficient s_j. The optimal choice is $s_j^* = C_{j,1}/C_{j,2}$, and the corresponding mean squared error is $C_{j,0} -$

$C_{j,1}^2/C_{j,2}$. These identities are adapted from Longford (1993), p. 14, for the case of a single rating session. The adjustment according to this scheme amounts to changing the score y_{ij} to $y_{ij} - s_j^* b_j$. This scheme is referred to as **sAdj**.

The true scores can be estimated without the intermediation of the estimates of severity. Consider the adjustment

$$\hat{\alpha}_i = y_{i,j_i} - u_i b_{j_i}, \tag{7}$$

where the coefficient u_i is determined so as to minimize the mean squared error $\mathbf{E}(\hat{\alpha}_i - \alpha_i)^2$. This adjustment scheme is described in Longford (1993), pp. 18–19. Using that notation, some simplification takes place when a single administration is used. For instance, $n_i^+ = n_{j_i}$, $n_i^- = 1/n_{j_i}$, and $R_i = 1/n_{j_i}$. The optimal shrinkage coefficient in (7) is $u_i^* = D_{i,1}/D_{i,2}$ and the attained minimum mean squared error is $D_{i,0} - D_{i,1}^2/D_{i,2}$, where

$$\begin{aligned} D_{i,0} &= \sigma_b^2 + \sigma_e^2 \\ D_{i,1} &= \sigma_b^2 \left(1 - \frac{n_{j_i}}{I}\right) + \sigma_e^2 \left(\frac{1}{n_{j_i}} - \frac{1}{I}\right) \\ D_{i,2} &= \sigma_a^2 \left(\frac{1}{n_{j_i}} - \frac{1}{I}\right) + \sigma_b^2 \left(1 - \frac{2n_{j_i}}{I} + \frac{n^{(2)}}{I}\right) + \sigma_e^2 \left(\frac{1}{n_{j_i}} - \frac{1}{I}\right). \end{aligned} \tag{8}$$

This adjustment scheme, referred to as **uAdj**, is similar to **sAdj**; compare the expressions in (6) and (8). Note in particular, that $C_{j_i,2} = D_{i,2}$. In essence, the scheme **uAdj** incorporates, in addition to severity, the information about inconsistency. The coefficients $s_{j_i}^*$ and $u_{j_i}^*$ differ substantially only when $\sigma_e^2(1/n_{j_i} - 1/I)$ is large relative to $C_{j,2}$, that is, when σ_e^2 is large (relative to σ_a^2 and σ_b^2), or when n_{j_i} is small.

Another adjustment scheme, referred to as **tAdj**, is based on shrinking toward the rater mean score. The score y_{i,j_i} is adjusted to

$$\hat{\alpha}_i = (1 - t_i) y_{i,j_i} + \frac{t_i}{n_{j_i}} \sum_{(i;\, j_i = j)} y_{i,j_i},$$

where the coefficient t_i is set so as to minimize the mean squared error $\mathbf{E}(\hat{\alpha}_i - \alpha_i)^2$. The optimal shrinkage coefficient is $t_i^* = E_{i,1}/E_{i,2}$ and the attained minimum mean squared error is $E_{i,0} - E_{i,1}^2/E_{i,2}$, where

$$
\begin{aligned}
E_{j,0} &= \sigma_b^2 + \sigma_e^2 \\
E_{j,1} &= \sigma_e^2 \left(1 - \frac{1}{n_{j,}}\right) \\
E_{j,2} &= \sigma_a^2 \left(1 - \frac{1}{n_{j,}}\right) + \sigma_e^2 \left(1 - \frac{1}{n_{j,}}\right),
\end{aligned} \tag{9}
$$

so that $t_i^* = \sigma_e^2/(\sigma_a^2 + \sigma_e^2)$. The fact that this coefficient does not depend on either σ_b^2 or n_j suggests that the scheme is deficient for some values of σ_b^2 and n_j. The scheme is of the same order of computational complexity as **uAdj**. Although its motivation is not as appealing as for the other two schemes, it performs better, in terms of the attained mean squared error, than the other two schemes, especially when the inconsistency variation is large.

Another adjustment scheme is based on the linear combination of the score, raters mean score, and the overall mean score. It is more efficient than either of the schemes considered because it is based on a wider class of linear functions. However, it is computationally much more demanding and less robust with respect to imprecison of the variances, and is therefore not considered here.

4 Analysis of the AP administrations

The following administrations of the AP tests were studied: AP Psychology, form 3PBP (6259 examinees), Computer Science, form 3PBP (6066 examinees), Computer Science, form A–3PBP (4414 examinees), and English Language and Composition, form 3PBP (35 689 examinees). Each test form administration took place in the fall 1992. All the essays were rated on the integer scale 0–9.

The Psychology test contained two essays (denoted A and B), the Computer Sciences test forms contained four essays each (denoted A, B, C, and D), and English Language and Composition contained three essays (A, B, and C).

In the AP tests, operationally, each essay is rated only once. In a reliability study a random sample of the essays is rerated once, by raters selected from the group of raters engaged in the operational scoring. The numbers of rerated essays were 299 for Psychology, 250 and 248 for the respective Computer Science

forms 3PBP and A-3PBP, and 250 for English Language and Composition. No reliability study was conducted for essay C of the English Language and Composition test.

The variance component estimates for these essays, based on the reliability study datasets, are listed in the middle portion of Table 2. The estimate of the inconsistency variance given by (3) was used in (5) for the analysis of the operational scores. The resulting estimates of the variance components due to the true scores ($\hat{\sigma}_a^2$) and rater severity ($\hat{\sigma}_b^2$) are given in the right-hand side of Table 2.

INSERT TABLE 2 HERE

The true score variances for the three English Language and Composition essays (1.2–2.0) are much smaller than those for the essays in any other test form (5.3–6.1 and one essay with variance 10.0). The inconsistency variances of the Computer Science essays are somewhat smaller than their counterparts in Psychology and English Language and Composition. No consistent pattern arises among the severity variances. Thus, the rater reliability, as measured by the correlation $\hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_b^2 + \hat{\sigma}_c^2)$, is by far the lowest for the English Language and Composition essays (0.50–0.73), followed by the Psychology essays (0.79–0.84), and the Computer Science essays (0.87–0.96). This finding suggests the natural hypothesis that grading is more reliable in sciences than in arts and humanities. The hypothesis could be tested by reanalyzing the reliability studies in other subject areas. Of course, the type of scoring rubric used may be a contributing factor to these differences.

For brevity, we give details of the analysis only for essays A and B in the Psychology test. The distributions of the unadjusted essay scores are depicted in Figure 1. The raters' workloads are listed in Table 3 together with the estimates of their severities. For both essay topics there are several raters who graded several hundred essays, but also several raters who graded only 1–20 essays. Note that some raters graded essays on both topics, although none of these raters graded an appreciable number of essays on both topics. In principle,

information about the raters could be pooled across the topics. However, this would be useful only if the raters graded large numbers of essays on both topics.

INSERT FIGURE 1 AND TABLE 3 HERE

The estimates of the rater severities are given in Table 3 in the rows labeled 'Severity'. For completeness, the shrinkage coefficients are given in the subsequent rows. For a fixed assignment design and a set of variance components, the shrinkage coefficient depends only on the rater's workload. The highest workload, for essay topic A, 657 essays, was assigned to rater 120. The shrinkage coefficient for his/her severity estimate is only 0.81. Thus, full adjustment (that is, no shrinkage) would have been far from optimal, in terms of the mean squared error, even for this rater. For essay topic B the shrinkage coefficients are somewhat higher because the severity variance $\hat{\sigma}_b^2$ is higher. For rater 106 who graded 726 essays, the shrinkage coefficient is equal to 0.97 which is very close to the full adjustment. Several raters with outlying severities can be identified in Table 3. For instance, rater 100 (essay B) has estimated severity −0.51 (based on 451 essays). The associated mean squared error is 0.14^2, and so there is ample evidence that this rater's severity is lower than, say −0.1.

The multiple-choice section of the AP Psychology test is formula scored. The observed scores are in the range 0–97.75, with mean 59 and median 60.75. The standard deviation of the scores is 16.2. The composite score is generated by adding the score from the multiple-choice part and the 2.7778-multiple of the total of the essay scores. The ranges of the scores for the multiple-choice parts of the other tests and the equations for combining the scores of the multiple-choice and essay components for the other analyzed administrations are given in Table 4. In addition, the cut-points for determining the AP grade are also given. For instance, the composite score of 112 or above is converted to grade 5, 94–111 to grade 4, and so on. Since the scores are not necessarily integers, the cut-points in this case are in fact 111.5, 93.5, and so on.

INSERT TABLE 4 HERE

Equations (6) and (8) indicate that the adjustment schemes **sAdj** and **uAdj**

are very similar, unless $\hat{\sigma}_e^2$ is a large fraction of $\hat{\sigma}_a^2 + \hat{\sigma}_b^2 + \hat{\sigma}_e^2$, or the workload n_j is small. Figure 2 contains plots of the adjustments for each essay in the Psychology test. Each rater is associated with an adjustment (a point in the plot); raters with workloads greater than 100 are denoted by black diamonds and those with workloads smaller than 100 are denoted by crosses. For orientation, the line representing identity is drawn in both plots. Clearly, the adjustment schemes differ noticeably only for raters with very small workloads. This affects only a small proportion of the examinees, though, whose score reliability is diminished by the absence of information about rater severity.

INSERT FIGURE 2 HERE

The impact of the score adjustment is most directly assessed by comparing the operational AP grades with the grades derived by using the adjusted scores in place of the operational scores for the free-response section of the test. The percentage of examinees that would be affected is given in Table 1. These percentages are very similar for the adjustment schemes **sAdj** and **uAdj**. Furthermore, there is a considerable overlap in the affected examinees. For instance, in AP Psychology the s-adjusted and u-adjusted AP grades differ only for eight examinees (0.13 per cent), by one point in each case. These are the only examinees whose s-adjusted AP grades agree with the operational AP grades but the u-adjusted grades do not. Each of these examinees was graded by at least one rater with a small workload.

The adjustment scheme **tAdj** shrinks the essay scores to rater means, and so it taps different information for improvement of the scores. The adjustment is of comparable quality with the schemes **sAdj** and **uAdj**, for raters with small workloads, but for large workloads it is clearly inferior. In particular, for English Language and Composition test the adjustment shrinks the AP grades radically towards the middle grade 3, thus bringing about a change for an unrealistically high percentage of the examinees. Of course, this shrinkage could be moderated by a change of the cut-points.

5 Conclusions

The study reported here demonstrates that the adjustment schemes described in Longford (1993), and summarized here in Section 3, can be implemented by combining information in the reliability studies with the operational grading in the AP tests. No alteration of the design of the reliability studies is necessary other than rescoring a random sample of essays for *every* essay topic. Once the inconsistency variance can be imputed, based on the estimates from a large number of previous forms of the same test, the reliability studies can be dispensed with. The proposed schemes are easy to implement and they involve no iterative procedures. The schemes **sAdj** and **uAdj** have a natural interpretation as adjustment for severity.

The schemes **sAdj** and **uAdj** differ only for essays graded by raters with small workloads, where the adjustment is associated with the greatest amount of uncertainty. When adjustment is applied the rating process could be improved by making sure there are no raters with small workloads. By extension, studies of rater adjustment based on subsets of examinees (due to limitations on the capacity to handle data) cannot make realistic conclusions about utility of score adjustment because they discard information about raters which has a strong impact on the efficiency of score adjustment.

6 Software

Software for the analyses described in this report was written in Splus, and translated to Fortran by Minhwei Wang. The input for the Fortran programs are the files containing examinee's records of the ratings and rater identification, together with the score from the multiple-choice section of the test. The equation for combining the scores and the cut-points for the AP grades have to be provided also. A different publication will contain detailed documentation of the software.

References

Braun, H.I. (1988) Understanding rater reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, **13**, 1–18.

Longford, N.T. (1993) Reliability of essay rating and score adjustment. ETS Technical Report 93-36, Educational Testing Service, Princeton, NJ.

Tables and Figures

Tables

Figures

Table 1: Percentages of examinees with altered grades as a result of adjusting the scores of the free-response sections.

AP test form		Adjustment schemes		
		uAdj	tAdj	sAdj
Psychology	Form 3PBP	2.75	7.21	2.62
Comp. Sci.	Form 3PBP	1.91	2.52	1.95
Comp. Sci.	Form A–3PBP	3.04	3.33	2.92
English L&C	Form 3PBP	6.38	42.23	6.31

Note: The adjustment schemes are described in Section 3.

Table 2: Variance component estimates in the reliability studies and in the operational rating.

Essay	Sample size		Reliability study			Operational	
			$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$
Psychology 3PBP							
A	299	6259	5.17	0.023	1.089	6.07	0.041
B	299	6259	4.55	0.339	1.315	5.28	0.113
Computer Science 3PBP							
A	250	6066	7.46	0.027	0.333	9.06	0.053
B	250	6066	8.89	0.000	0.702	10.04	0.096
C	250	6066	7.80	0.000	0.488	8.03	0.064
D	250	6066	6.50	0.000	0.248	5.63	0.027
Computer Science A–3PBP							
A	248	4414	5.15	0.010	0.522	5.62	0.301
B	248	4414	6.38	0.186	0.411	6.39	0.284
C	248	4414	5.89	0.000	0.464	6.04	0.204
D	248	4414	4.75	0.000	0.429	5.39	0.114
English L&C 3PBP							
A	250	35689	1.07	0.311	1.129	1.23	0.123
B	250	35689	1.80	0.133	0.619	2.03	0.146
C	0	35689			0.874	1.07	0.141

Notes: A reliability study for Essay C in the English Language and Composition test was not conducted. The estimate of the inconsistency variance, 0.874, is the mean of the estimates for the Essays A and B.

Table 3: Workloads and severity estimates of the essay raters in the AP Psychology test.

Raters' workloads and severity estimates												
Essay A												
Rater Id.	102	103	104	106	107	108	109	111	112	113	114	115
Workload	1	368	350	25	439	442	1	339	574	551	538	580
Severity	−0.01	−0.08	0.13	0.03	−0.07	−0.03	0.01	0.06	−0.21	−0.25	0.19	0.09
Shrinkage	0.01	0.67	0.66	0.12	0.71	0.71	0.01	0.69	0.77	0.76	0.76	0.78
$\sqrt{\text{MSE}}$	0.20	0.12	0.12	0.19	0.12	0.12	0.20	0.12	0.11	0.11	0.11	0.11
Rater Id.	117	120	122	124	180	181	182	183	187	190	999	
Workload	2	657	552	582	9	6	72	59	1	37	14	
Severity	−0.03	−0.03	0.20	0.16	−0.04	−0.03	−0.13	−0.11	0.01	−0.03	−0.19	
Shrinkage	0.01	0.81	0.76	0.77	0.05	0.03	0.29	0.25	0.01	0.17	0.07	
$\sqrt{\text{MSE}}$	0.20	0.11	0.11	0.11	0.20	0.20	0.17	0.18	0.20	0.18	0.19	
Essay B												
Rater Id.	100	101	102	106	107	108	109	110	116	117	118	119
Workload	451	405	564	726	4	1	535	495	295	413	321	353
Severity	−0.51	0.03	−0.02	0.11	0.09	0.02	−0.19	0.10	−0.06	0.17	0.24	−0.06
Shrinkage	0.88	0.87	0.92	0.97	0.06	0.02	0.91	0.90	0.82	0.87	0.83	0.84
$\sqrt{\text{MSE}}$	0.14	0.15	0.14	0.13	0.33	0.33	0.14	0.14	0.16	0.15	0.15	0.15
Rater Id.	120	121	123	125	152	180	181	182	183	190	196	999
Workload	3	498	484	433	1	125	115	1	3	21	1	12
Severity	−0.19	−0.01	0.37	0.12	0.05	0.16	−0.45	−0.08	−0.01	−0.40	−0.03	−0.81
Shrinkage	0.05	0.90	0.89	0.88	0.02	0.66	0.64	0.02	0.05	0.26	0.02	0.17
$\sqrt{\text{MSE}}$	0.33	0.14	0.14	0.14	0.33	0.20	0.20	0.33	0.33	0.29	0.33	0.31

Notes: Rater '999' is an amalgam of raters used to resolve contingency cases. 'Workload' is the number of essays graded by the rater. 'Shrinkage' is the shrinkage coefficient applied for the estimate of rater severity. $\sqrt{\text{MSE}}$ is the square root of the mean squared error of the estimator of the rater's severity.

Table 4: Information about combining the scores from the multiple-choice and free-response parts of the AP tests.

Test	Multiple choice scores	Composite score	Cut-points
Psychology	0–97.75	M + 2.7778 F	50.5, 73.5, 93.5, 111.5
Comp. Sci	0–40	0.9722 M + 0.8750 F	21.5. 32.5, 43.5, 51.5
Comp. Sci A	0–40	1.3889 M + 1.2500 F	33.5, 49.5, 63.5, 70.5
English L&C	0–53	1.2736 M + 3.0556 F	48.5, 74.5, 91.5, 107.5

Notes: The second column gives the ranges of the scores for the multiple-choice part of the test. The third column is the formula for combining multiple-choice scores (M) with the scores from the essay part (free-response part, F). The right-most column gives the cut-points for the AP grades 1–5.

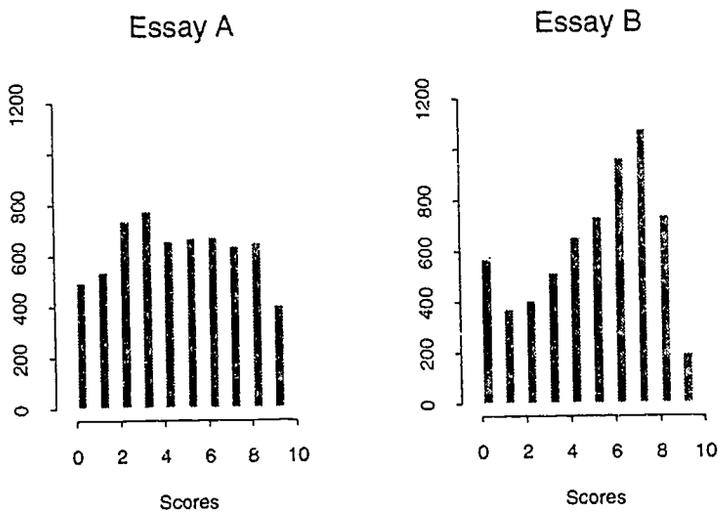

Figure 1: Distributions of unadjusted essay scores and estimates or rater severity in the AP Psychology test.

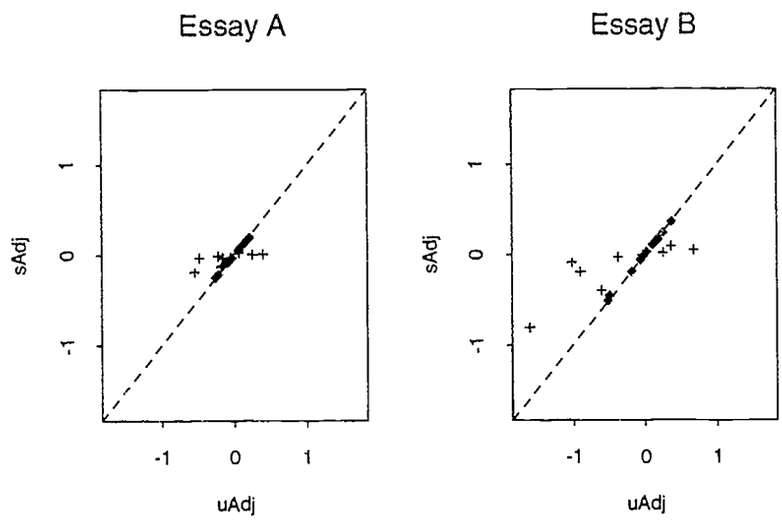

Figure 2: Adjustments for severity in the AP Psychology test.
The horizontal axis represents adjustment **uAdj**, the vertical axis adjustment **sAdj**. The raters who graded fewer than 100 essays are marked by a crosses, those who graded more than 100 essays are marked by a black diamonds.